Classification multi-labels graduée: Apprendre les relations entre les labels ou limiter la propagation d'erreur ?

Khalil Laghmari*,**, Christophe Marsala**, Mohammed Ramdani*

*Laboratoire Informatique de Mohammedia,
FSTM, Hassan II University of Casablanca,
BP 146 Mohammedia 20650 Maroc
laghmari.khalil@gmail.com, ramdani@fstm.ac.ma
**Sorbonne Universités,
UPMC Univ Paris 06,
CNRS, LIP6 UMR 7606,
4 place Jussieu 75005 Paris, France
christophe.marsala@lip6.fr

Résumé. La classification multi-labels graduée est la tâche d'affecter à chaque donnée l'ensemble des labels qui lui correspondent selon une échelle graduelle de degrés d'appartenance. Les labels peuvent donc avoir à la fois des relations d'ordre et de co-occurrence.

D'un côté, le fait d'ignorer les relations entre les labels risque d'aboutir à des prédictions incohérentes, et d'un autre côté, le fait de prendre en compte ces relations risque de propager l'erreur de prédiction d'un label à tous les labels qui lui sont reliés.

Les approches de l'état d'art permettent soit d'ignorer les relations entre les labels, soit d'apprendre uniquement les relations correspondant à une structure de dépendance figée. L'approche que nous proposons permet l'apprentissage des relations entre les labels sans fixer une structure de dépendance au préalable. Elle est basée sur un ensemble de classifieurs mono-labels, un pour chaque label. L'idée est d'apprendre d'abord toutes les relations entre les labels y compris les relations cycliques. Ensuite les dépendances cycliques sont résolues en supprimant les relations d'intérêt minimal. Des mesures sont proposées pour évaluer l'intérêt d'apprendre chaque relation. Ces mesures permettent d'agir sur le compromis entre l'apprentissage de relations pour une prédiction cohérente et la minimisation du risque de la propagation d'erreur de prédiction.

1 Introduction

La classification multi-labels (Tsoumakas et Katakis (2007) ; Zhang et Zhou (2014)) consiste à affecter à chaque donnée un ou plusieurs labels en même temps, et la classification ordinale (Frank et Hall (2001)) consiste à affecter à chaque donnée un label

selon une échelle graduelle de degrés d'appartenance. La classification multi-labels graduée (CMLG) (Cheng et al. (2010)) est donc considérée comme une combinaison de la classification multi-labels et de la classification ordinale, où chaque donnée est associée à un ensemble de labels selon une échelle graduelle de degrés d'appartenance. La CMLG peut être aussi considérée comme un cas particulier de la classification floue (Bouchon-Meunier et al. (1997)), où l'intervalle $[0, 1]$ est remplacé par un ensemble fini et ordonné de valeurs exprimant des degrés d'appartenance.

Les défis relevés par la CMLG que traite ce papier sont, premièrement de trouver la bonne structure de dépendance entre les labels sans l'imposer au préalable, et deuxièmement de considérer le compromis entre prendre compte des relations entre les labels afin d'obtenir des prédictions cohérentes, et ne pas les prendre en compte afin de limiter le risque de propagation d'erreur.

La Section 2 introduit une formalisation du problème de la CMLG, et discute des approches de l'état d'art permettant de le résoudre partiellement. La Section 3 présente les principes de l'approche que nous proposons, et illustre son déroulement sur un jeu de données synthétisées. La Section 4 discute les résultats obtenus sur des jeux de données réelles. La Section 5 conclut ce travail et discute des pistes pour les travaux futurs.

2 Etat de l'art

Soit $X = \{x_i\}_{1 \leq i \leq n}$ l'ensemble des données, $C = \{c_l\}_{1 \leq l \leq k}$ l'ensemble des labels, et $M = \{m_g\}_{1 \leq g \leq s}$ l'ensemble ordonné des degrés d'appartenance. Chaque donnée x_i est un vecteur de valeurs $(x_{ij})_{1 \leq j \leq p}$. A chaque donnée x_i correspond un vecteur y_i de degrés d'appartenance $(y_{il})_{1 \leq l \leq k}$.

x_i est dit vecteur de valeurs des attributs descriptifs, et y_i est dit vecteur de valeurs des attributs de décision. L'ensemble des valeurs que peut prendre le l-ème attribut de décision est noté M_l. L'espace $M_1 \times \ldots \times M_k$ est noté M_*^k. L'objectif de la CMLG est d'apprendre un classifieur $H : X \rightarrow M_*^k$ permettant de prédire pour chaque donnée $x_i \in X$ le vecteur de degrés d'appartenance correspondant $y'_i = H(x_i)$.

Un classifieur H adapté à la CMLG peut être construit à partir d'un ensemble de classifieurs, un par label ou un par paire de labels. La base des approches apprenant un classifieur par label est l'approche *Binary Relevance* (BR) dont l'inconvénient est de ne pas tenir compte les relations entre les labels. Les approches visant à remédier à ce problème telles que *Classifier Chains* (CC) (Read et al. (2011)) et *Classifier Treillis* (Read et al. (2015)) ont l'inconvénient d'imposer une structure de dépendance entre les labels, et de ne permettre l'apprentissage que pour les relations respectant cette structure (Laghmari et al. (2015)). Les approches basées sur un classifieur pour chaque paire de labels (Hüllermeier et al. (2008) ; Fürnkranz et al. (2008)) ont l'avantage d'apprendre les relations de préférences entre chaque paire de labels. Les approches *Horizontal Calibrated Label Ranking (Horizontal CLR), Full Calibrated Label Ranking (Full CLR), et Joined Calibrated Label Ranking* (Joined CLR) (Brinker et al. (2014)) sont basées sur l'apprentissage de préférences entre les labels, et ont l'inconvénient de constuire plus de classifieurs que les approches verticales.

3 L'approche proposée : PSI-MC

La première idée du méta-classifieur proposé (PSI-MC) est d'apprendre un **ensemble initial de classifieurs** $H^0 = \{H_l\}_{1 \leq l \leq k}$, un pour chaque label en considérant les autres labels en tant qu'attributs descriptifs (Laghmari et al. (2016)). Ceci permet d'apprendre des relations entre les labels sans fixer une structure de dépendance, mais peut conduire éventuellement à des dépendances cycliques entre les classifieurs. La deuxième idée de PSI-MC est d'apprendre un **ensemble final de classifieurs** $\mathbb{H} = \{\mathbb{H}_l\}_{1 \leq l \leq k}$ à partir de H^0 en supprimant les dépendances cycliques entre les classifieurs.

Le tableau 1 illustre un ensemble de données d'apprentissage ayant deux attributs descriptifs $\{a_1, a_2\}$, cinq labels $\{c_1, c_2, c_3, c_4, c_5\}$, et quatre degrés d'appartenance possibles $\{0, 1, 2, 3\}$. La figure 1 illustre les classifieurs initiaux obtenus par l'algorithme $J48$ de Weka (Hall et al. (2009)). L'étape 0 dans la figure 2 illustre les relations de dépendance entre les classifieurs de H^0. Un arc allant d'un nœud H_l vers un nœud $H_{l'}$ représente le fait que $H_{l'}$ dépend de H_l. Nous considérons la fonction $D^{\rightarrow} : H \rightarrow \mathcal{P}(H)$ qui pour chaque classifieur fournit l'ensemble de classifieurs qui dépendent de lui, et la fonction $D^{\leftarrow} : H \rightarrow \mathcal{P}(H)$ qui pour chaque classifieur fournit l'ensemble de classifieurs dont il dépend.

Une **mesure de présélection** $\mathbb{P} : H \rightarrow \{0, 1\}$ décide si un classifieur a besoin d'être remplacé : $\mathbb{P}(H_l) = 1$, ou non : $\mathbb{P}(H_l) = 0$. Dans l'exemple étudié, la mesure $\mathbb{P}$ présélectionne les classifieurs qui dépendent d'au moins un autre classifieur. Elle est définie telle que : $\forall H_l \in H : \mathbb{P}(H_l) = 1$ si $|D^{\leftarrow}(H_l)| \geq 1$, et $\mathbb{P}(H_l) = 0$ sinon. Les classifieurs non présélectionnés sont indépendants et donc directement ajoutés à l'ensemble $\mathbb{H}$ (Étape 1 de la figure 2).

Une **mesure de sélection** $\mathbb{S} : \mathcal{P}(H) \rightarrow H$, choisit un classifieur $H_L \in \{H_l \in H, P(H_l) = 1\}$ à remplacer dans $\mathbb{H}$ par un nouveau classifieur H'_L. Dans l'exemple étudié, la mesure $\mathbb{S}$ sélectionne le premier classifieur dont dépend le plus de classifieurs. Elle est définie par $\mathbb{S}(H) = \underset{H_l \in H}{argmax}(|D^{\rightarrow}(H_l)|)$. En utilisant cette mesure à l'étape 1 de la figure 2, le classifieur H_2 est sélectionné.

	a_1	a_2	c_1	c_2	c_3	c_4	c_5
x_1	20	20	0	0	3	0	0
x_2	30	40	1	0	3	0	0
x_3	20	30	0	0	3	0	0
x_4	20	10	0	0	0	0	3
x_5	50	40	2	3	0	1	2
x_6	50	20	2	3	0	1	2
x_7	10	10	0	1	2	2	3
x_8	10	30	0	3	1	2	2
x_9	10	10	0	1	2	2	3
x_{10}	10	50	0	3	1	2	2

TAB. 1: Données synthétisées

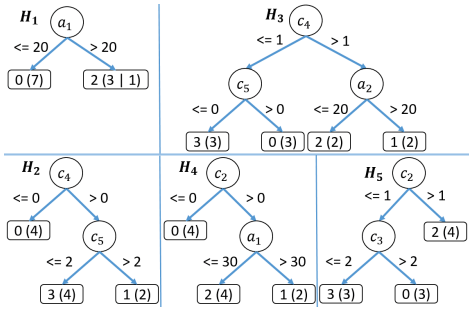

FIG. 1: Arbres de décision de l'ensemble H^0

Une **mesure d'intérêt de chaînage** des classifieurs $\mathbb{I} : \mathbb{H} \rightarrow \{0, 1\}$ décide pour chaque classifieur H'_L s'il peut considèrer en tant qu'attribut descriptif le label corres-

Apprendre les relations entre les labels ou limiter la propagation d'erreur ?

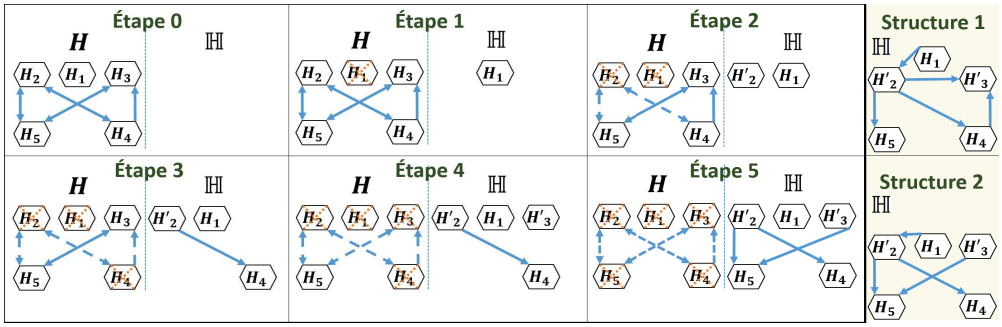

FIG. 2: Déroulement des étapes du méta-classifieur proposé : PSI-MC

pondant à un classifieur final $\mathbb{H}_l$. Dans l'exemple étudié, la mesure $\mathbb{I}$ ne permet aucun chaînage. Elle est définie telle que : $\forall \mathbb{H}_l \in \mathbb{H} : \mathbb{I}(\mathbb{H}_l) = 0$. En utilisant cette mesure à l'étape 1 de la figure 2 le classifieur H'_2 est appris en ne considérant aucun attribut descriptif suplémentaire. H'_2 est donc indépendant (étape 2 de la figure 2).

Après avoir éliminé le classifieur H_2 de H^0, le classifieur H_4 devient indépendant et il est donc ajouté directement à $\mathbb{H}$. Le lien de dépendance reliant H_4 à H_2 est remplacé par un lien le reliant à H'_2 (étape 3 de la figure 2).

La procédure d'élimination des classifieurs de H^0 et d'ajout des classifieurs dans $\mathbb{H}$ est répétée jusqu'à ce que H^0 devient vide (étape 5 de la figure 2). L'ensemble final de classifieurs obtenu $\mathbb{H}$ ne présente aucune dépendance cyclique. Il est possible de trouver d'autres structures de dépendance en utilisant par exemple, une mesure d'intérêt de chaînage autorisant tous les chaînages : $\forall \mathbb{H}_l \in \mathbb{H} : \mathbb{I}(\mathbb{H}_l) = 1$ (structure 1 de la figure 2), ou une mesure autorisant le chaînage uniquement avec des classifieurs indépendants : $\forall \mathbb{H}_l \in \mathbb{H} : \mathbb{I}(\mathbb{H}_l) = 1$ *si* $|D^{\leftarrow}(H_l)| = 0$ (structure 2 de la figure 2).4

4 Expérimentation

Afin d'évaluer l'erreur de prédiction sur un ensemble de données non classées $X = \{x_i\}_{1 \leq i \leq n}$, nous utilisons une mesure étendue de la *distance de Hamming* au cas de la CMLG, définie telle que : $\forall i \in [1, n]$: $\textit{hamming-loss}(x_i) = \sum_{1 \leq l \leq k} \frac{|\mathbb{H}_l(x_i) - y_{il}|}{k \times |m_s - m_1|}$

et $\textit{hamming-loss}(X) = \sum_{1 \leq i \leq n} \frac{\textit{hamming-loss}(x_i)}{n}$

L'ensemble de données initiales *BelaE*[1] est construit en demandant à 1930 étudiants d'age et de sexe différents, d'indiquer l'importance de 48 propriétés (labels) de leurs points de vue selon une échelle de 1 à 5.

Afin de remédier au problème du manque d'attributs descriptifs, seulement k propriétés sont considérées en tant qu'attributs de décision. Les attributs restants sont considérés en tant qu'attributs descriptifs.

50 jeux de données sont générés à partir du jeu de données BelaE en variant à chaque fois les attributs de décision. Les 50 jeux de données sont fournis pour le cas

1. http ://www.ke.tu-darmstadt.de/resources/CMLG

$k = 5$ (BelaE-5) et pour le cas $k = 10$ (BelaE-10). Ceci permet de se comparer aux approches *Full CLR*, *Joined CLR*, et *Horizontal CLR* (Brinker et al. (2014)) (tableau 2 et tableau 3).

Données	méta-classifieur	Moyenne de l'erreur de Hamming et déviation standard
BelaE-5	Horizontal CLR	0.1577 ± 1.53
	Joined CLR	0.1796 ± 1.31
	PSI-MC ($\mathbb{I} = 0$)	0.1889 ± 0.0949
	PSI-MC ($\mathbb{I} = 1$)	0.1891 ± 0.0956
	Full CLR	0.3397 ± 5.79

TAB. 2: Résultats obtenus sur BelaE-5

Données	méta-classifieur	Moyenne de l'erreur de Hamming et déviation standard
BelaE-10	Horizontal CLR	0.1513 ± 0.95
	Joined CLR	0.1792 ± 0.87
	PSI-MC ($\mathbb{I} = 0$)	0.1884 ± 0.0709
	PSI-MC ($\mathbb{I} = 1$)	0.1894 ± 0.0721
	Full CLR	0.3544 ± 3.70

TAB. 3: Résultats obtenus sur BelaE-10

Les deux configurations du méta-classifieur proposé PSI-MC sont les plus stables selon l'écart type de hamming-loss. De plus. Elles permettent l'obtention de prédictions plus précises que l'approche *Full CLR*, et presque aussi précise que l'approche *Joined CLR*.

La configuration $\mathbb{I} = 0$ permet un apprentissage plus rapide puisqu'elle ignore les relations entre les labels au remplacement d'un classifieur, et la configuration $\mathbb{I} = 1$ fournit plus d'information et plus d'interprétabilité puisqu'elle permet l'apprentissage d'un maximum de relations entre les labels.

5 Conclusion et perspectives

Certaines approches de l'état d'art ne considèrent pas les relations entre les labels, d'autres imposent une structure de dépendance dès le départ. Dans ce travail nous proposons un méta-classifieur PSI-MC qui permet à la fois de trouver une bonne structure de dépendance sans l'imposer au début, et de gérer le compromis entre cohérence de prédiction et limitation de la propagation d'erreur. Les premiers résultats obtenus montrent que le méta-classifieur proposé fournit des résultats de prédiction comparables et plus stables par rapport aux approches auxquelles il a été comparé. Pour les travaux futurs, nous envisageons de faire plus d'expérimentations sur différents jeux de données. Nous envisageons aussi d'étudier l'impact du changement des mesures du méta-classifieur proposé sur les résultats de prédiction.

Références

Bouchon-Meunier, B., C. Marsala, et M. Ramdani (1997). *Learning from Imperfect Data*, pp. 139–148. John Wiley & Sons.

Brinker, C., E. L. Mencía, et J. Fürnkranz (2014). Graded multilabel classification by pairwise comparisons. In *2014 IEEE International Conference on Data Mining*, pp. 731–736.

Cheng, W., K. Dembczynski, et E. Hüllermeier. (2010). Graded multilabel classification : The ordinal case. In M. Atzmüller, D. Benz, A. Hotho, et G. Stumme (Eds.),

Proceedings of LWA2010 - Workshop-Woche : Lernen, Wissen & Adaptivitaet, Kassel, Germany.

Frank, E. et M. Hall (2001). A simple approach to ordinal classification. In *Proceedings of the 12th European Conference on Machine Learning*, EMCL '01, London, UK, UK, pp. 145–156. Springer-Verlag.

Fürnkranz, J., E. Hüllermeier, E. Loza Mencía, et K. Brinker (2008). Multilabel classification via calibrated label ranking. *Machine Learning 73*(2), 133–153.

Hall, M., E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, et I. H. Witten (2009). The weka data mining software : An update. *SIGKDD Explor. Newsl. 11*(1), 10–18.

Hüllermeier, E., J. Fürnkranz, W. Cheng, et K. Brinker (2008). Label ranking by learning pairwise preferences. *Artificial Intelligence 172*(16–17), 1897 – 1916.

Laghmari, K., C. Marsala, et M. Ramdani (2016). Graded multi-label classification : Compromise between handling label relations and limiting error propagation. In *2016 11th International Conference on Intelligent Systems : Theories and Applications (SITA)*, pp. 1–6.

Laghmari, K., M. Ramdani, et C. Marsala (2015). A distributed graph based approach for rough classifications considering dominance relations between overlapping classes. In *SITA'15, Intelligent Systems Theories and Applications, 2015 10th Inte. Conf. on*, pp. 1–6.

Read, J., L. Martino, P. M. Olmos, et D. Luengo (2015). Scalable multi-output label prediction : From classifier chains to classifier trellises. *Pattern Recognition 48*(6), 2096 – 2109.

Read, J., B. Pfahringer, G. Holmes, et E. Frank (2011). Classifier chains for multi-label classification. *Mach. Learn. 85*(3), 333–359.

Tsoumakas, G. et I. Katakis (2007). Multi-label classification : An overview. *Int J Data Warehousing and Mining 2007*, 1–13.

Zhang, M. L. et Z. H. Zhou (2014). A review on multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering 26*(8), 1819–1837.

Summary

Graded multi-label classification is the task of associating to each data a set of labels according to an ordinal scale of membership degrees. Therefore labels can have both order and co-occurrence relations. On the one hand, ignoring label relations may lead to inconsistent predictions, and on the other hand, considering those relations may spread the prediction error of a label to all related labels.

Unlike state of art approaches which can learn only relations fitting a predefined dependency structure, our proposed approach doesn't set any predefined structure. The idea is to learn all possible relations, then resolve cyclic dependencies using appropriate measures. Those measures allow managing the compromise between considering label relations for a consistent prediction, and ignoring them to minimize the prediction error propagation.